Genetics and Molecular Biology, 24 (1-4), 1-7 (2001)

The libraries that made SUCEST

André L. Vettore[1,a], Felipe R. da Silva[1,b], Edson L. Kemper[1,c] and Paulo Arruda[14]

Abstract

A large-scale sequencing of sugarcane expressed sequence tags (ESTs) was carried out as a first step in depicting the genome of this important tropical crop. Twenty-six unidirectional cDNA libraries were constructed from a variety of tissues sampled from thirteen different sugarcane cultivars. A total of 291,689 cDNA clones were sequenced in their 5' and 3'end regions. After trimming low-quality sequences and removing vector and ribosomal RNA sequences, 237,954 ESTs potentially derived from protein-encoding messenger RNA (mRNA) remained. The average insert size in all libraries was estimated to be 1,250bp with the insert length varying from 500 to 5,000 bp. Clustering the 237,954 sugarcane ESTs resulted in 43,141clusters, from which 38% had no matches with existing sequences in the public databases. Around 53% of the clusters were formed by ESTs expressed in at least two libraries while 47% of the clusters are formed by ESTs expressed in only one library. A global analysis of the ESTs indicated that around 33% contain cDNA clones with full-length insert.

INTRODUCTION

Single-pass sequencing of cDNAs to generate "expressed sequence tags" (ESTs) has proven to be a powerful, economical and rapid approach to identify genes that are preferentially expressed in certain tissue or cell types of multicellular organisms (Adams *et al.*, 1991, Hwang *et al.*, 1997, Liew *et al.*, 1994, Adams *et al.*, 1995). Increasing importance has also been attributed to ESTs as a tool for the annotation of complete genome sequences of mammalians and plants. Unique ESTs provided biological evidence of hundreds of predicted genes, newly discovered genes, or transcript isoforms leading to considerable advance in gene identification mission in multicellular organisms (Andrews *et al.*, 2000). Today, more than ten million ESTs are currently available through the dbEST entry of GenBank (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_sumary.html); however, only 14% of dbEST release 022301 of February 23, 2001 corresponds to plant sequences.

Another useful aspect of ESTs is in accessing genetic information of species with a complex genome, whose access is difficult using conventional genetics. This is the case of sugarcane, an important crop that is cultivated in the tropics for its high sucrose accumulation in the stalk. Among the cultivated crops, sugarcane possesses perhaps one of the most complex genomes (for a review see Grivet and Arruda, 2002). Modern sugarcane cultivars are hybrids derived from the crossing of *Saccharum officinarum*, usually having 2n = 80 chromosomes and *Saccharum spontaneum*, 2n = 40 - 128 chromosomes. In view of the structural differences between chromosomes of the two species, the hybrids possess different proportions of chromosomes, varying chromosome sets and complex recombinational events (Grivet and Arruda, 2002). This imposes tremendous difficulties in applying conventional plant breeding techniques to sugarcane.

As a first step in depicting the sugarcane genome, the ONSA consortium (Simpson and Perez 1998) launched in September of 1998 the Sugarcane Expressed Sequence Tag project (SUCEST), aiming at sequencing random ESTs and identifying around 50,000 unique genes (http://sucest.lad.ic.unicamp.br/en/).

To improve the probability of getting a maximum number of different ESTs, researchers have been using normalized and/or subtracted cDNA libraries that bring the frequency of each clone in a cDNA library within a narrow range (Soares and Bonaldo 2000). However, normalization and/or subtraction procedures are in general laborious and have the tendency of increasing the proportion of small insert clones. In the SUCEST project we have implemented an efficient procedure to generate conventional cDNA libraries to generate large scale ESTs from sugarcane. This paper describes the construction of these libraries, representing all major organs, harvested at different developmental stages and used to generate one of the largest plant EST collections.

[1]*Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, C.P. 6109, 13083-970 Campinas, SP, Brazil and Depto de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, C.P. 6109, 13083-970 Campinas, SP, Brazil.*
Present address:
[a]*Instituto Ludwig de Pesquisa sobre o Câncer, 01509-010, São Paulo, SP, Brazil.*
[b]*Embrapa Agrobiologia, BR 465, Km 47, CxP 74505, 23851-970, Seropédica, RJ, Brazil.*
[c]*Monsanto Company, GG5B, 700 Chesterfield Pkwy N, 63038, Chesterfield - MO, USA.*
Send correspondence to Paulo Arruda. E-mail: parruda@unicamp.br.

MATERIAL AND METHODS

Plant material

Sugarcane tissues were obtained from commercial cultivars (Table I) grown at the Copersucar experimental station (Piracicaba, SP, Brazil), at the Universidade Federal de São Carlos experimental station (Serra do Ouro, AL, Brazil) and at the Centro de Biologia Molecular e Engenharia Genética (Campinas, SP, Brazil). After harvesting, tissues were frozen in liquid nitrogen and stored at -80 °C.

RNA isolation

Total RNA was isolated using Trizol (Invitrogen) according to manufacturer's instructions. Due to the high carbohydrate content and the presence of phenolic compounds, total RNA from immature seeds was isolated according to the method described by Manning (1991).

Poly(A)$^+$ mRNA was purified from total RNA using Oligotex-dT (Qiagen) according to manufacturer's instructions. Purity and RNA integrity were assessed by absorbance at 260/280 nm and agarose gel electrophoresis.

cDNA library construction

Libraries were constructed using the SuperScript cDNA Synthesis and Plasmid Cloning Kit (Invitrogen) according to the manufacturer's protocols. One microgram of poly(A) + mRNA was reverse-transcribed using a poly-dT primer containing the *Not*I site. The efficiency of cDNA synthesis was monitored with radioactive nucleotides. The second cDNA strand was then synthesized by replacing the RNA in the hybrids with DNA by using a combination of RNase H, DNA Polymerase I and DNA Ligase. After the second-strand synthesis and ligation of *Sal*I adapters, cDNA was digested with *Not*I, generating cDNA flanked by *Sal*I sites at 5' ends and *Not*I sites at the 3' ends. Excess adapters were removed and cDNAs were size fractioned in a 40 cm long 1 mm ID Sepharose CL-2B columm. One hundred and fifty µL fractions were collected and 8 µL aliquots of each fraction was electrophoresed in 1.5%

Table I - Description of the SUCEST Libraries.

Library code	Library name	Description	Sugarcane variety
AD1	*G. diazotroficans1*	Mixture of tissues from root to shoot zone, stem and apical meristem of plantlets cultivated *in vitro* and infected with *Gluconacetobacter diazotroficans*	P70-1143[3]
AM1, AM2	Apical Meristem	Apical meristem of young plants	SP80-3280[2]
CL6	Calli	Pool of *calli* treated for 12 h at 4 °C and 37 °C in the dark or ligth	SP80-3280[1]
FL1, FL3, FL4, FL5, FL8	Flower 1, 3, 4, 5 and 8	Flowers harvested at different developmental stages	SP80-87432[1] PB5211 x P57150-4[1]
HR1	*H. rubrisubalbicans1*	Mixture of tissues from root to shoot zone, stem and apical meristem of plantlets cultivated *in vitro* and infected with *Herbaspirilum diazotroficans*	SP70-1143[3]
LB1, LB2	Lateral Bud 1 and 2	Lateral buds from mature plants	SP80-3280[1]
LR1, LR2	Leaf Roll 1 and 2	Leaf roll from immature plants	SP80-3280[1]
LV1	Leaf 1	Etiolated leaves from plantlets grown *in vitro*	SP83-5077 SP80-185 SP87-396 SP80-3280 SP803280 x SP81-5441[1]
RT1, RT2, RT3	Root 1, 2 and 3	0.3 cm-length roots from mature plants and root apex	SP80-3280[1]
RZ1, RZ2	Root to shoot	Root to shoot zone of young plants	SP80-3280[1]
RZ3	zone 1, 2 and 3		
SB1	Stalk Bark 1	Stalk bark from mature sugarcane plants	SP80-3280[2]
SD1, SD2	Seeds 1 and 2	Developing seeds	CB47-89 RB855205 RB845298 RB805028[4]
ST1, ST3	Stem 1 and 3	First and fourth internodes of immature plants	SP80-3280[1]

cDNA libraries were constructed from different tissues sampled from different varieties grown at Copersucar experimental station (Piracicaba-SP)[1], CBMEG - Universidade Estadual de Campinas (Campinas-SP)[2], Universidade Federal do Rio de Janeiro (Rio de Janeiro-RJ)[3], and Universidade Federal de São Carlos experimental station (Serra do Ouro-AL)[4].

agarose gel to determine the size range of cDNAs. Fractions with cDNAs with a minimum size of 500 base pair (bp) were pooled and ligated to pSPORT1 vector (Invitrogen) predigested with *Sal*I and *Not*I. The resulting plasmids were transformed in DH10B cells (Invitrogen) by electroporation. Unamplified libraries were plated and individual colonies picked and transferred to 96 well plates containing liquid Circle Grow (CG) medium (BIO 101), supplemented with 100 mg/L of ampicillin and 8% glycerol. Three copies of each cDNA clone were stored at -80 °C.

Template preparation and DNA sequencing

DNA template preparations and sequencing reactions were performed in a 96-well format. Plasmid templates were prepared using modified alkaline lysis (http://sucest.lad.ic.unicamp.br). Sequencing reactions were performed on plasmid templates using a quarter of the standard volume of ABI Prism BigDye Terminator Sequencing Kit (Applied Biosystems) and the T7 promoter primer (5'-TAATACGACTCACTATAGGG-3') that hybridizes upstream of the *Sal*I site in the pSPORT1 polylinker (5'end of the cDNA inserts) or the SP6 promoter primer (5'-ATTTAGGTGACACTATAG-3') that hybridizes downstream of the *Not*I site (3'end of the cDNA inserts). Reaction products were precipitated with 95% ethanol using sodium acetate (3M) and Glycogen (1g/L) as carriers and washed twice with 75% ethanol before drying under vacuum. The sequencing reaction products were analyzed on 377-96 ABI Sequencers.

Sequence analysis

Sequencing of sugarcane ESTs was performed by 23 laboratories located in Universities and Research Institutes of the State of São Paulo and sequences were processed by the Bioinformatics laboratory (LBI) located at Instituto de Computação, Universidade Estadual de Campinas. A detailed description of the methods used to receive, process, analyze, and display the sequences along with additional tools to help explore the sequence data can be found in this issue (Telles *et al.*, 2001, Telles and da Silva, 2001).

RESULTS AND DISCUSSION

The SUCEST strategy

EST programs to acquire information about the transcriptome has been carried out for hundreds of organisms including plants and mammals. In most of the cases unidirectionally cloned cDNA libraries have been prepared using bacterial or phage vectors, so that the 5' and/or 3' end of the clones can be sequenced. Since single pass reads result in average ~350 high quality nucleotides, sequencing 3' ends covers mainly the untranslated region of the transcript. Moreover, the 3'end of the cDNA clones contain a long poly-A tail that is useless in terms of biological information and in general introduces technical difficulties in the sequencing process. However, because the untranslated 3'end represent the less conserved region of the transcripts it is useful, for example, to avoid misassembly of reads coming from highly conserved sequences from members of gene families. Sequencing 5' ends of unidirectional cDNA clones, on the other hand, can be of great advantage for large scale EST projects. Since the 5' untranslated region is shorter, it is likely that it contains protein-coding sequences. In addition, because a large proportion of clones present partial cDNA sequences, it is possible to collect enough information to assemble the full consensus sequence of a transcript, increasing the likelihood that database searches will result in the assignment of their putative functions. Based on this assumptions we decide sequence the 5' end of the cDNA clones to build up the SUCEST database.

The libraries

Table I shows the description of the libraries used in the SUCEST project. A variety of tissues were sampled from different cultivars, in order to access transcript information of genes expressed in many biological systems. Two libraries AD1 and HR1 were constructed using tissues from *in vitro* cultured plantlets infected with *Gluconacetobacter diazotroficans* and *Herbaspirilum diazotroficans.* These are endophytic nitrogen fixing bacteria that colonize sugarcane tissues (Lee, *et al.*, 2000). Sequencing from these libraries could lead to discovery of genes involved in plant-bacteria interaction and in nitrogen assimilation in sugarcane. Libraries AM1, AM2, LB1 and LB2 were constructed using apical meristem of young plants and lateral buds from adult plants. These libraries shall contribute with genes expressed at the initial stages of organ differentiation. Calli produced from sugarcane meristems was used in an experiment devised to access genes induced by cold and heat. Two weeks old calli was incubated at 4 °C or 37 °C for 12 h. Part of the tissues was maintained in the dark and part in continuous light. The CL6 library was prepared with a mixture of equal amounts of RNA extracted from these tissues and it is expected that this library will contribute with genes induced by cold and heat. FL1, FL3, FL4, FL5 and FL8 are libraries constructed from flower tissues harvested at different developmental stages and may contribute with genes expressed in this important plant organ. To access information on genes expressed in leaves, we constructed LR1 and LR2 libraries from leaf roll of adult plants and LV1 from etiolated leaves of plantlets grown *in vitro*. A collection of libraries representing roots or tissues from which roots emerge are represented by RT1, RT2 and RT3 which are libraries constructed from roots

sampled from plantlets grown *in vitro* or plants grown in greenhouse, while RZ1, RZ2 and RZ3 were constructed from root to shoot zone of young plants grown in greenhouse. SB1 is a library constructed from stalk bark of adult plants and may contribute with genes involved in the synthesis of cell wall components including waxes. SD1 and SD2 are libraries constructed from developing seeds. Finally, we constructed the libraries ST1 and ST3 from first and fourth internodes of adult plants at the time of intense sucrose synthesis and accumulation.

Quality control

Large-scale sequencing demands care with the quality of biological materials and accurate performance at each step of the process, both to provide sequence data of the highest possible quality and to detect or avoid mistakes (Adams *et al.*, 1995). At each step of the SUCEST project, from tissues sampling to sequence analysis, quality control and evaluation procedures were used to assess the accuracy of the data. The goal of the SUCEST project was that cDNA libraries should contain all sequences present in the initial poly(A)$^+$ mRNA population, which is useful to access expression profile through electronic Northern; unidirectionally cloned so that the orientation of each cDNA is known, facilitating subsequent sequence analysis; include a large proportion of full-length inserts; and reveal low levels of contamination with genomic or ribosomal RNA. Table II shows the quality control steps used during cDNA library construction and sequencing. Tissues were quickly frozen in liquid nitrogen, RNA quality analyzed by different methods and the cDNAs were synthesized and size selected using special gel filtration columns. cDNAs were unidirectionally cloned in pSPORT plasmidial vector and introduced into DH10B competent cells. Libraries with title less than 1 x 10^4 were discarded. Colonies were placed into 96 well plates and stored at -80 °C. A sample of ~400 clones from each library was examined to evaluate library quality, such as percentage of clones with no inserts, percentage of ESTs with exact matches to sequences derived from ribosomal RNA species, *E. coli* or bacteriophage lambda, percentage of ESTs with no significant matches to any sequence in the public databases, and an estimate of the number of clusters that contain a full-length coding region sequence. Libraries selected for EST analysis typically exhibited a broad diversity of transcripts (no single gene or small group of genes dominating the distribution), a low percentage of clones with no insert, a low percentage of ribosomal RNA clones, and no evidence of contamination with sequences from other organisms. The libraries that did not meet these general criteria were discarded.

Sequencing in the SUCEST project was carried out using ABI377 sequencers, which are prone to error during gel tracking. To minimize errors the 8th row of each 96 well plates was used to build control plates that were resequenced. Computer analysis was then used to check the address match. These allowed the SUCEST project to keep the address error to less than 5%, so that a sequence in the computer corresponds, with high fidelity, to a clone in the freezer.

Table II - Quality control and evaluation of SUCEST libraries.

Parameter	Quality control and evaluation
Tissue sampling	Tissues snap frozen quickly after harvesting
Poly(A)$^+$ RNA purification	Purity and RNA integrity were assessed by absorbance at 260/280 nm and agarose gel electrophoresis
cDNA synthesis	Tracer levels of ^{32}P used; agarose gel examination for degradation; column chromatography for size selection
cDNA library construction	Blue/white screen for inserts; PCR to check insert sizes; libraries must contain at least 10^5 recombinants
Library storage	All clones were grown in 96 well plates containing CG media supplemented with 8% glycerol > Plates were stored at -80 °C in triplicate
Sample sequencing	Around 400 clones of each library were sequenced to check gene diversity, contaminations and rRNA
Clone address	One clone in each twelve was resequenced to detect putative address mistakes
Template preparation	DNA quality and concentration checked by agarose gels

Quality control procedures for each step in the EST process are listed with specific points of evaluation or standards to be met.

SUCEST data set

Table III shows the summary of the complete data set of the SUCEST project. A total of 259,325 cDNA clones were sequenced in their 5'end region and 32,364 of them had also their 3'end region sequenced. Therefore, the project produced 291,689 ESTs. After trimming of low-quality sequences and removal of vector and ribosomal RNA ESTs, 237,954 ESTs potentially derived from protein-encoding messenger RNA (mRNA) remained. This represents a success index of 81.56%, which is comparable with other EST projects worldwide. Before entering the sequencing pipeline, each SUCEST cDNA library was evaluated for the average size of cDNA inserts. cDNA libraries that contained an average insert size below 500bp were discarded. The average insert size in all libraries was estimated to be 1,250bp (n = 4,000) (Table III). The distribution of the insert length was between 500 and 5,000bp. In order to clone genes encoding low molecular weight proteins, we constructed some cDNA libraries (LR2, RZ2 and SD2 - See Table IV) with an average insert size of 855bp.

Table III - Summary of SUCEST data.

Analyzed data	
Total ESTs	291,689
5' ESTs	259,325
3' ESTs	32,364
ESTs remaining after trimming quality control	237,954
Average insert size, bp	1,250
Average EST length, bp	750
Average EST bases with Phred quality ≥ 20	365

Numbers of sequenced cDNA clones and generated ESTs from 26 libraries constructed from different sugarcane tissues. 259,325 ESTs were generated by sequencing the 5' end of cDNA clones. Another 32,364 ESTs were generated by sequencing the 3' end of cDNAs clones. The average insert size was calculated for 400 cDNA clones from each library. The EST length and the number of bases with Phred quality ≥ 20 was calculated from the total EST set.

Table IV - Characteristics of the SUCEST libraries.

Library code	Average insert size (bp)	Sequenced clones	Valid reads	Success index (%)	Novelty (%)
AD1	1,330	18,144	14,701	81.02	55.34
AM1	1,300	12,480	10,881	87.19	55.05
AM2	-	15,648	13,403	85.65	49.45
CL6	1,150	7,392	5,518	74.65	63.62
FL1	1,400	18,528	15,343	82.81	54.82
FL3	1,340	13,056	10,727	82.16	53.26
FL4	1,370	16,896	13,964	82.65	52.19
FL5	1,180	10,080	7,744	76.83	66.05
FL8	1,400	5,184	4,652	89.74	72.26
HR1	-	12,000	9,729	81.08	52.11
LB1	1,150	7,488	5,879	78.51	62.91
LB2	1,660	10,560	8,953	84.78	60.33
LR1	1,240	14,112	11,701	82.92	56.85
LR2	870	4,128	3,418	82.80	68.13
LV1	1,260	6,432	4,557	70.85	67.32
RT1	1,450	8,640	7,255	83.97	58.26
RT2	1,400	12,288	10,606	86.31	54.86
RT3	1,000	10,560	7,441	70.46	58.54
RZ1	1,290	3,168	2,831	89.36	71.07
RZ2	-	5,760	5,031	87.34	63.14
RZ3	-	15,168	12,862	84.80	50.75
SB1	-	16,320	13,189	80.81	56.16
SD1	1,240	11,040	8,601	77.91	51.84
SD2	840	10,368	8,505	82.03	48.19
ST1	1,050	8,448	6,933	82.07	62.87
ST3	1,350	12,000	8,939	74.49	50.55

The average insert size of each library was determined in a sample of 400 clones by gel electrophoresis of the clones digested with PvuII. Valid reads are defined as reads containing at least 140 bp with Phred quality ≥ 20. The success index is the number of valid reads in relation to the number of clones sequenced. The Novelty represents the probability of a new sequence to be founding in the library.

After the trimming process, all new sequences were compared to the previous sequences that had already been deposited in the SUCEST database. Every time that an EST was similar to a sequence that already existed in the database, both were grouped together in a cluster. As noted in Table V, the 237,954 valid sequences were assembled into 43,141clusters.

Each cluster consensus sequence was compared against the non-redundant nucleotide and peptide databases (GenBank) using the programs BLASTN and BLASTX. Sequences that did not match these databases were further compared against the dbEST. Using a blast *E*-Value threshold (Altschul *et al.*, 1997) equal to or below e-5, of the 43,141 SUCEST clusters, 26,525 (61.5%) had matches with an existing sequence in GenBank (Table V). Therefore, 16,616 (38.5%) of the SUCEST clusters could potentially represent new genes. These values are comparable to those found for ESTs sequences from other organisms (Hwang *et al.* 2000; Adams *et al.* 1992; Claverie 1996). Ascribing functions to those anonymous sequences has therefore become one of the major bottlenecks in plant and animal genomics.

Tissue and cellular differentiation depend on specific patterns of gene expression. Therefore, in large-scale EST sequencing, sampling many different tissues and in different physiological conditions increases the chance to pick up transcripts rare in one cell type but less rare in another. SUCEST database was built up with sequences derived from 26 libraries constructed from different tissues sampled at different developmental stages (Table I) and an average of 10,000 clones were sequenced from each library. Sequencing from many libraries resulted in a novelty ratio as good as the ratios found in other EST projects that used normalized libraries (Bonaldo *et al.*, 1996).

Around 53.2% of the SUCEST clusters were formed by ESTs expressed in at least two libraries. This suggests that these genes are being coordinately expressed in different tissues or that they are expressed in response to specific physiological conditions or developmental requirements. On the other hand, 46.8% of the clusters (Table VI - the sum of specific contributions) are formed by ESTs expressed in only one library. This suggests that these ESTs could correspond to genes expressed in a tissue/time fashion, varying in different tissue/physiological conditions. Nonetheless, these data should be analyzed taking into account that 16,338 (37.9%) are singletons, therefore representing rare transcripts. The uniformity in the amount of singletons in the different libraries (Table VI) strengthens the value of the approach adopted.

A global analysis of all SUCEST clusters indicated that around 33% contain cDNA clones with full-length in-

Table V - Statistics of EST clustering and contiging.

ESTs analyzed	237,954
Total clusters (C+S)	43,141
Clusters with at least 2 reads (C)	26,303
Singletons (S)	16,838
C+S sequences finding homolog in GenBank	26,525
C+S sequences with no homolog in GenBank	16,616
C+S with full length insert	14,409

ESTs were clustered using CAP3 assembler (Huang and Madan, 1999). The *E*-value cut of threshold to be considered for C or S as having homology to other proteins in the nr GenBank database using BLASTX was ($<10^{-5}$). Clones were considered as having a putative full length insert when their sequences started within the first 15 amino acids of their hit in GenBank. C or S were considered as having tentative full consensus sequence when their sequences started within the first 15 amino acids and finished within the last 15 amino acid of their hit in GenBank.

Table VI - EST clustering in the individual libraries.

Library code	Number of clusters	Unique clusters	Number of singletons	Specific contribution (%)
AD1	10,736	3,120	2,821	3.84
AM1	7,870	1,930	1,726	2.37
AM2	9,079	2,389	2,012	2.94
CL6	4,282	1,231	1,112	1.51
FL1	11,438	3,740	3,468	4.60
FL3	7,847	2,178	1,997	2.68
FL4	10,145	2,626	2,407	3.23
FL5	6,489	1,697	1,589	2.08
FL8	3,963	811	780	0.99
HR1	6,697	1,664	1,434	2.04
LB1	4,697	1,149	1,074	1.41
LB2	7,056	1,749	1,597	2.15
LR1	8,867	2,250	2,104	2.77
LR2	2,901	696	662	0.85
LV1	4,005	1,037	950	1.27
RT1	5,706	1,435	1,336	1.76
RT2	7,851	2,081	1,875	2.56
RT3	5,699	1,398	1,251	1.72
RZ1	2,374	448	426	0.55
RZ2	4,054	939	869	1.15
RZ3	8,858	2,331	2,094	2.86
SB1	10,204	2,910	2,774	3.58
SD1	6,114	1,600	1,451	1.96
SD2	5,696	1,856	1,539	2.28
ST1	5,682	1,431	1,341	1.76
ST3	6,124	1,335	1,253	1.64

The number of clusters that contain one or more reads from a specific library is indicated, as well as, the clusters that were formed only by reads of a specific library (Unique Clusters). The number of clusters that were formed by only one read (Singleton) is also indicated. The specific contribution is calculated dividing the Unique Clusters of each library by the total number of clusters (43,141).

serts (Table V). This is in accordance with the results obtained in the mouse EST project (Marra *et al.*, 1999).

This collection of 237,954 ESTs provides us with a preliminary view into the gene expression profile of sugarcane. The identification of genes involved in different cellular processes suggests that the generation of large-scale ESTs should provide valuable insights into the molecular mechanisms of plant function and development.

REFRENCES

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C. and **Venter, J.C.** (1992). Sequence identification of 2,375 human brain genes [see comments]. *Nature 355* (6361): 632-634.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B. and **Moreno, R.F.** (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science 252* (5013): 1651-1656.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D. and **White, O.** (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature 377* (6547 Suppl): 3-174.

Altschul, S.F., Madden, T.L., Schiffer, A.A., Zhang, J., Zhang, Z., Miller, W. and **Lipman, D.J.** (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res 25* (17): 3389-3402.

Andrews, J., Bouffard, G.G., Cheadle, C., Lü, J., Becker, K.G. and **Oliver, B.** (2000). Gene Discovery Using Computational and Microarray Analysis of Transcription in the *Drosophila melanogaster* Testis. *Genome Res 10*: 2030-2043.

Bonaldo, M.F., Lennon, G. and **Soares, M.B.** (1996). Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res 6* (9): 791-806.

Claverie, J.M. (1996). Exploring the vast territory of uncharted ESTs. In: *Genomes, molecular biology and drug discovery*. Academic Press, pp. 56-71.

Green, P. (1999). phrap.doc: http://bozeman.genome.washington.edu/phrap.docs/phrap.html

Grivet, L. and **Arruda, P.** (2002). Sugarcane genomics: depicting the complex genome of an important tropical crop. *Current Opinion in Plant Biology 5*: 122-127.

Huang, X. and **Madan, A.** (1999). CAP3: A DNA sequence assembly program. *Genome Res. 9*: 868-877.

Hwang, D.M., Dempsey, A.A., Lee, C.Y. and **Liew, C.C.** (2000). Identification of differentially expressed genes in cardiac hypertrophy by analysis of expressed sequence tags. *Genomics 66* (1): 1-14.

Hwang, D.M., Dempsey, A.A., Wang, R.X., Rezvani, M., Barrans, J.D., Dai, K.S., Wang, H.Y., Ma, H., Cukerman, E., Liu, Y.Q., Gu, J.R., Zhang, J.H., Tsui, S.K., Waye, M.M., Fung, K.P., Lee, C.Y. and **Liew, C.C.** (1997). A genome-based resource for molecular cardiovascular medicine: toward a compendium of cardiovascular genes. *Circulation 96* (12): 4146-4203.

Lee, S., Reth, A., Meletzus, D., Sevilla, M. and **Kennedy, C.** (2000) Characterization of a Major Cluster of nif, fix, and

Associated Genes in a Sugarcane Endophyte, *Gluconacetobacter diazotrophicus. Journal of Bacteriology 182*: 7088-7091.

Liew, C.C., Hwang, D.M., Fung, Y.W., Laurenssen, C., Cukerman, E., Tsui, S. and **Lee, C.Y.** (1994). A catalogue of genes in the cardiovascular system as identified by expressed sequence tags. *Proc Natl Acad Sci U.S.A. 91* (22): 10645-10649.

Manning, K. (1991). Isolation of nucleic acids from plants by differential solvent precipitation. *Anal Biochem 195* (1): 45-50.

Marra, M., Hillier, L., Kucaba, T., Allen, M., Barstead, R., Beck, C., Blistain, A., Bonaldo, M., Bowers, Y., Bowles, L., Cardenas, M., Chamberlain, A., Chappell, J., Clifton, S., Favello, A., Geisel, S., Gibbons, M., Harvey, N., Hill, F., Jackson, Y., Kohn, S., Lennon, G., Mardis, E., Martin, J. and **Waterston, R.** (1999). An encyclopedia of mouse genes. *Nat Genet 21* (2): 191-194.

Soares, M. and **Bonaldo, M.** (2000). Constructing and screening normalized cDNA libraries. In: *Genome analysis: a laboratory manual* (Birren, B., Green, E., Klapholz, S., Myers, R. and Roskams, A., eds.), CSHL Press: CSHL Press, pp. 49-157.

Simpson, A.J.G. and **Perez, J.F.** (1998). Latin America - ONSA, the Sao Paulo virtual genomics institute. *Nat Biotechnol 16*: 795-796.

Telles, G.P., Braga, M.D.V., Dias, Z., Lin, T., Quitzau, J.A.A., da Silva, F.R. and **Meidanis, J.** (2001). Bioinformatics of the sugarcane EST project. *Genetics and Molecular Biology 24* (1-4): 9-15.

Telles, G.P. and **da Silva, F.R.** (2002). Trimming and clustering sugarcane ESTs. *Genetics and Molecular Biology 24* (1-4): 17-23.